Proceedings of Machine Learning Research vol 245:1–7, 2024 2024 International Conference on Machine Learning and Intelligent Computing

City Gas Load Forecasting Based on PCCs-CNN-LSTM Model

Guangjun Zai ZAIGUANGJUN@ZZU.EDU.CN
School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou, Henan, 450001, China

Yuwei Zhang 1257266813@QQ.COM
School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou, Henan, 450001, China

Sen Wu 124665802@QQ.COM
School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou, Henan, 450001, China

Zhao Tian* TIANZHAO@ZZU.EDU.CN
School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou, Henan, 450001, China

Editors: Nianyin Zeng and Ram Bilas Pachori

Abstract

The forecast of urban gas load is of great significance for the safety and stability of gas supply, to ensure the normal production activities of residents. The influence factors of sunshine duration were introduced, and the nine identified influencing factors were analyzed by Pearson correlation coefficient (PCCs). According to the correlation, the optimal input was selected one by one. The influencing factors with high correlation were used as the input of Convolutional Neural Networks (CNN) and Long short-term memory (LSTM), respectively, to forecast the daily load, monthly load and quarterly load of urban gas, and verify their accuracy and effectiveness. The results show that the optimal number of input factors for daily load forecasting and monthly load forecasting is 5, and the optimal number of input factors for quarterly load forecasting is 8. For daily load forecasting, the absolute percentage errors of monthly load forecasting and quarterly load forecasting of PCS-CNN-LSTM model are 3.94%, 4.61% and 5.73% respectively. The root mean square error and mean absolute error of PCS-CNN-LSTM model are better than that of a single LSTM model.

Keywords: gas load forecasting, convolutional neural networks, LSTM neural network, Pearson correlation coefficient

1. Introduction

In recent years, with China's increasingly stringent requirements for environmental pollution control, the proportion of natural gas as a clean energy source in China's primary energy has increased year by year (Shen et al., 2023), and China's natural gas consumption has shown a rapid growth trend, which has put forward higher requirements for the efficient use and rational allocation of natural gas resources (Gorucu, 2004a). Ensuring the safety and stability of residents' gas demand (Aras and Aras, 2004) is conducive to ensuring people's livelihood. Gas load prediction (Yucesan et al., 2021) is of positive significance for the efficient and reasonable allocation of urban natural gas resources and the safety assessment of gas pipeline network, and is conducive to the scientific scheduling of gas system and the reduction of energy loss (Potočnik and Govekar, 2016).

2. PCCS-CNN-LSTM Models

To improve the accuracy of gas load prediction, this paper proposes a combined model based on PCCs-CNN-LSTM, which integrates PCCs, CNN, and LSTM network models (Laib et al., 2019), into aprediction framework (Figure 1). Based on the engineering experience, nine factors were preliminarily selected from the perspectives of historical load, meteorological factors, date type, etc.: weather conditions, maximum temperature, minimum temperature, average temperature, wind level, air quality, heating season, holiday level, and sunshine duration.

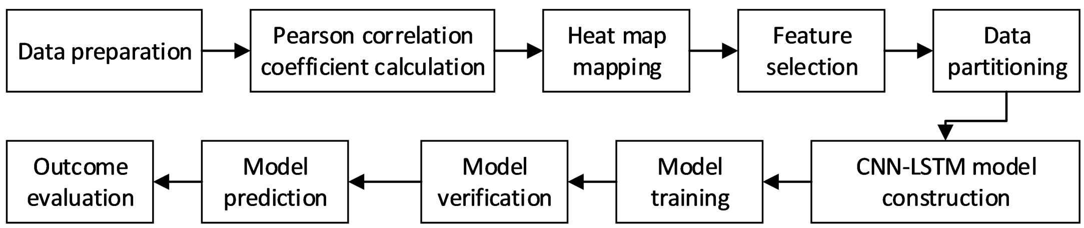

Figure 1: PCCs-CNN-LSTM workflow Diagram.

The Pearson correlation coefficient method was used to compare the correlations, eliminate the factors with very weak correlations, and pre-process the collected data. The CNN network contains convolutional computation and has a deep structure, which adopts the method of weight sharing and local joining, which can directly obtain valid features from the original data and automatically extract the local features of the data (Ni et al., 2019). Since CNNs are not sensitive to differences in time series, but to spatial features, the result may not be ideal if only CNNs are used to predict temperatures. Although the LSTM network is sensitive to time series, it does not have the same strong image data acquisition ability as CNN, and due to its huge nonlinear modeling ability, the generalization ability of model prediction is insufficient. If the LSTM neural network is used alone to predict temperatures, overfitting can occur. In this paper, we will use a prediction method based on CNN and LSTM, which first uses the powerful feature extraction ability of CNN to extract the features, and then input the features into the LTSM to further integrate and predict the result.

2.1. Pearson Correlation Coefficient

Due to the different degrees of influence of different factors on gas load, if the influencing factors with a small correlation degree are taken as the input of the model, the prediction accuracy of the model and the complexity of the network will be affected (Gorucu, 2004b). Therefore, PCCs is widely used to measure the correlation between two variables in the field of natural sciences, especially in the analysis of complex data with the comprehensive influence of multiple factors and has high reliability. In Pearson correlation analysis, the expectation, variance, and covariance are estimated based on the sample, and the Pearson correlation coefficient is defined as the quotient of the covariance and standard deviation of the estimation sample (Eq.(1)).

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \quad (1)$$

where X, Y are the mean of n test values. The value of r is between -1 and +1, and the greater the absolute value of r, the stronger the correlation. The Pearson correlation coefficient between each feature is calculated, the correlation coefficient matrix is generated, the heat map is drawn according to the correlation coefficient matrix to visualize the correlation between each feature, and the influence degree of each influencing factor on the natural gas load can be obtained according to the correlation size between each feature. To determine the number of influencing factors that make the prediction accuracy reach the optimum, the factors are eliminated one by one according to the correlation degree from low to high, the remaining factors are substituted into the prediction model, and the optimal number of influencing factors is determined according to the prediction results.

2.2. CNN-LSTM Model

CNN can be used to extract effective features and information from gas load data, and then combine these features and information with LSTM neural networks that can memorize long-term information in the field of gas load forecasting, which can more accurately predict the fluctuation of gas load (Xie et al., 2020; Lang et al., 2019).

Figure 2 shows the network structure of CNN-LSTM neural network gas load prediction model (Zha et al., 2022), which includes convolutional layer, pool layer, Dropout layer, LSTM and fully connected layer.

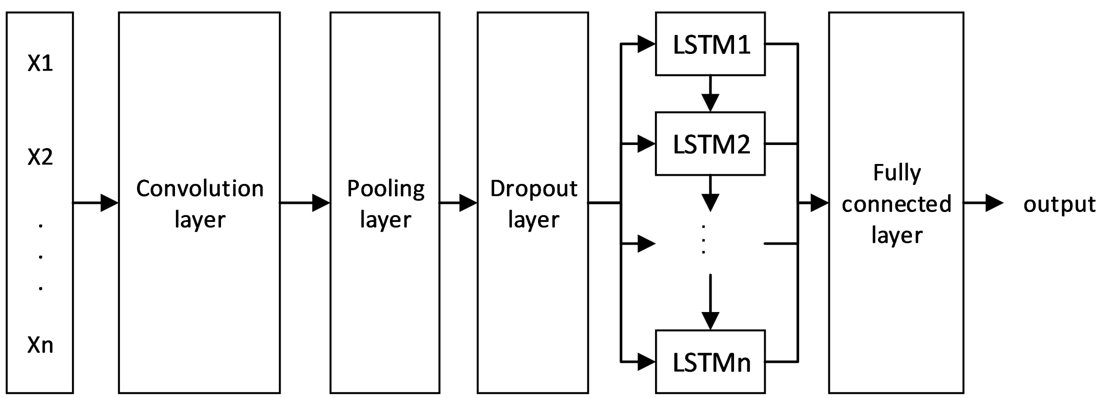

Figure 2: CNN-LSTM network structure.

The basic idea is as follows: in the first step, the gas load data and its influencing factor data are taken as inputs, and the salient features are extracted through the convolutional layer to obtain the local correlation between the influencing factors; In the second step, the number of parameters is compressed through the pooling layer, which can further reduce the dimensionality of the data. In the third step, by adding a fully connected layer with a dropout mechanism, the conger avoids the phenomenon of overfitting as much as possible by randomly discarding the connections of some neurons with a certain probability. The fourth step is to use LSTM for prediction; In the fifth step, the data is converted into a one-dimensional structure through the fully connected layer, so as to obtain the final output result.

3. Results

3.1. Dataset

This study collects data on the overall gas load of a city from January 2017 to October 2022, as well as daily temperature data, weather conditions, wind data, sunshine hours, and air quality data on the weather data website. According to the actual situation of the city, the heating season and holiday levels are divided.

Since the input of the machine learning model requires numerical data, the two factors of weather type and date type on that day are quantified (Table 1 and Table 2), and the sample database of natural gas load and quantified influencing factors is established.

Table 1: Weather Condition Quantification.

Sunny	Cloudy	Haze	Fog	Overcast	Rain	Snow	Other
1	2	3	4	5	6	7	8

Table 2: Date Types Quantification.

Workday	Holiday
0	1

3.2. Model Evaluation Metrics

The daily load pre-measurement, monthly load pre-measurement and seasonal load pre-measurement are divided into three types. To comprehensively evaluate the forecasting accuracy of the PCS-CNN-LSTM daily load forecasting model, the following evaluation indexes are selected.

1. Mean Absolute Error (MAE) (Eq. (2)).

$$\boldsymbol{MAE = \frac{1}{N}\sum_{f=1}^{N} \mid y_f - \hat{y}_f \mid} \tag{2}$$

2. Root Mean Squared Error (RMSE) (Eq. (3)).

$$\boldsymbol{RMAE = \sqrt{\frac{1}{N}\sum_{f=1}^{N} (y_f - \hat{y}_f)^2}} \tag{3}$$

3. Mean Absolute Percentage Error (MAPE) (Eq. (4)).

$$\boldsymbol{MPAE = \frac{1}{N}\sum_{f=1}^{N} \left(\frac{\mid y_f - \hat{y}_f \mid}{y_f}\right) \times 100\%} \tag{4}$$

MAPE is a measure of the overall error of a model. The smaller the values of these evaluation indicators, the higher the prediction accuracy of the model.

3.3. PCCs-CNN-LSTM Performance Analysis

The combined model of PCS-CNN-LSTM was used to make daily, monthly and quarterly prediction of the city. The number of different influencing factors was selected and sorted according to correlation. $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9$ were used to represent Heating season, Average temperature, minimum temperature, maximum temperature, sunshine duration, air quality, wind rating, holiday rating, weather conditions. Table 3, Table 4 and Table 5 reflect the impact of influencing factors on daily forecast, monthly forecast and quarterly forecast respectively.

Table 3: Daily forecast of influencing factors.

Influencing factor	MAE	RMSE	MAPE
$X_1X_2X_3X_4$	118402	177442	4.02%
$X_1X_2X_3X_4X_5$	113590	166271	3.94%
$X_1X_2X_3X_4X_5X_6$	122524	182811	4.23%
$X_1X_2X_3X_4X_5X_6X_7$	145496	211689	5.08%
$X_1X_2X_3X_4X_5X_6X_7X_8$	146019	221198	4.92%
$X_1X_2X_3X_4X_5X_6X_7X_8X_9$	138821	204566	4.66%

Table 4: Monthly forecast of influencing factors.

Influencing factor	MAE	RMSE	MAPE
$X_1X_2X_3X_4$	162148	232755	6.77%
$X_1X_2X_3X_4X_5$	111785	175908	4.61%
$X_1X_2X_3X_4X_5X_6$	170511	275335	7.05%
$X_1X_2X_3X_4X_5X_6X_7$	164978	228396	7.04%
$X_1X_2X_3X_4X_5X_6X_7X_8$	165727	228396	7.02%
$X_1X_2X_3X_4X_5X_6X_7X_8X_9$	144981	241399	6.04%

Table 5: Quarterly forecast of influencing factors.

Influencing factor	MAE	RMSE	MAPE
$X_1X_2X_3X_4$	201088	253966	8.57%
$X_1X_2X_3X_4X_5$	147722	200093	6.15%
$X_1X_2X_3X_4X_5X_6$	158652	207708	6.75%
$X_1X_2X_3X_4X_5X_6X_7$	165302	212120	7.20%
$X_1X_2X_3X_4X_5X_6X_7X_8$	138277	195496	5.73%
$X_1X_2X_3X_4X_5X_6X_7X_8X_9$	143910	199454	6.05%

To sum up, by comparing the data obtained in the table without quantitative influencing factors, it can be concluded that the optimal results are obtained when choosing five influencing factors for daily and monthly prediction of natural gas load. For the seasonal prediction of natural gas load, 8 influencing factors were selected as the best results.

Table 6, Table 7, Table 8 and Figure 3 reflect the daily, monthly and quarterly forecast results of PCS-LSTM and PCS-CNN-LSTM models under different influencing factors.

Table 6: Daily forecast result.

Model	MAE	RMSE	MAPE
LSTM	132104	209356	4.47%
CNN-LSTM	138821	204566	4.66%
PCCs-CNN-LSTM	113590	166271	3.94%

Table 7: Monthly forecast result.

Model	MAE	RMSE	MAPE
LSTM	263835	404091	7.89%
CNN-LSTM	144981	241399	6.04%
PCCs-CNN-LSTM	111785	175908	4.61%

Table 8: Quarterly forecast result.

Model	MAE	RMSE	MAPE
LSTM	189014	262798	7.86%
CNN-LSTM	143910	199454	6.05%
PCCs-CNN-LSTM	138277	195496	5.73%

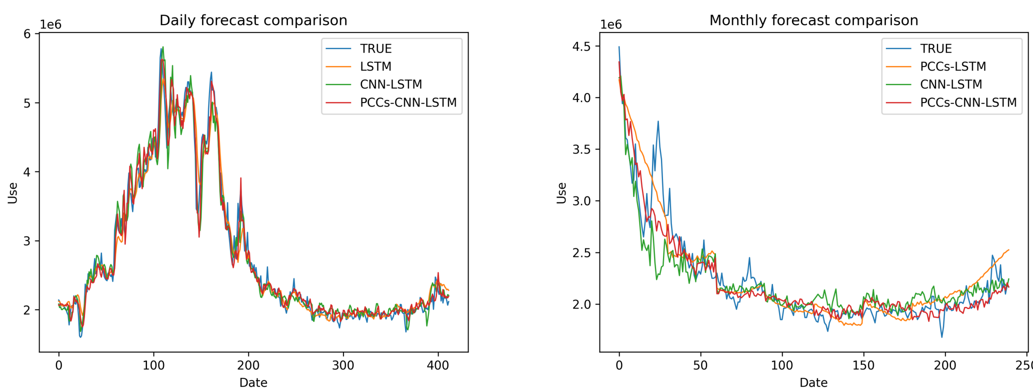

Figure 3: Gas forecast result of day and month.

In summary, for the urban load data, comparing the performance data of the chart and table, it is found that under the same parameters, the PCS-CNN-LSTM model after screening the influencing factors performs better on all three indicators (MAE, RMSE, MAPE). This shows that the model is more accurate in predicting gas loads in cities.

4. Conclusion

By employing Pearson correlation coefficients, this study clearly shows that the accuracy of the data can be improved when considering the duration of insolation in daily, monthly, and quarterly forecasts. In terms of natural gas load forecasting, the performance of the PCCs-CNN-LSTM model is significantly better than that of a single LSTM model. This shows a significant advantage in improving prediction accuracy when considering models that integrate PCCs, CNNs, and LSTMs.

References

Haydar Aras and Nil Aras. Forecasting residential natural gas demand. *Energy Sources*, 26(5): 463–472, 2004.

Fatma B Gorucu. Evaluation and forecasting of gas consumption by statistical analysis. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 26(3):267–276, 2004a.

Fatma B Gorucu. Evaluation and forecasting of gas consumption by statistical analysis. *Energy Sources*, 26(3):267–276, 2004b.

Oussama Laib, Mohamed Tarek Khadir, and Lyudmila Mihaylova. Toward efficient energy systems based on natural gas consumption prediction with lstm recurrent neural networks. *Energy*, 177: 530–542, 2019.

Kun Lang, Mingyuan Zhang, Yongbo Yuan, and Xijian Yue. Short-term load forecasting based on multivariate time series prediction and weighted neural network with random weights and kernels. *Cluster Computing*, 22:12589–12597, 2019.

Lina Ni, Yujie Li, Xiao Wang, Jinquan Zhang, Jiguo Yu, and Chengming Qi. Forecasting of forex time series data based on deep learning. *Procedia computer science*, 147:647–652, 2019.

Primož Potočnik and Edvard Govekar. Applied short-term forecasting for the slovenian natural gas market. In *2016 13th International Conference on the European Energy Market (EEM)*, pages 1–5, 2016.

Xin Shen, Shuhui Zhou, and Chunhua Wu. Development of china's natural gas market in 2022 and outlook for 2023. *International Petroleum Economics*, 31(4):25–36, 2023.

Hailun Xie, Li Zhang, and Chee Peng Lim. Evolving cnn-lstm models for time series prediction using enhanced grey wolf optimizer. *IEEE access*, 8:161519–161541, 2020.

Melih Yucesan, Engin Pekel, Erkan Celik, Muhammet Gul, and Faruk Serin. Forecasting daily natural gas consumption with regression, time series and machine learning based methods. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 21(7):1–16, 2021.

Wenshu Zha, Yuping Liu, Yujin Wan, Ruilan Luo, Daolun Li, Shan Yang, and Yanmei Xu. Forecasting monthly gas field production based on the cnn-lstm model. *Energy*, 260:124889, 2022.